Temporal Knowledge Graph Question Answering: A Survey

Anonymous ACL submission

Abstract

Knowledge Base Question Answering (KBQA) has been a long-standing field to answer questions based on knowledge bases. Recently, the evolving dynamics of knowledge have attracted a growing interest in Temporal Knowledge Graph Question Answering (TKGQA), an emerging task to answer temporal questions. However, this field grapples with ambiguities in defining temporal questions and lacks a systematic categorization of existing methods for TKGQA. In response, this paper provides a thorough survey from two perspectives: the taxonomy of temporal questions and the methodological categorization for TKGQA. Specifically, we first establish a detailed taxonomy of temporal questions engaged in prior studies. Subsequently, we provide a comprehensive review of TKGQA techniques of two categories: semantic parsing-based and TKG embedding-based. Building on this review, the paper outlines potential research directions aimed at advancing the field of TKGQA. This work aims to serve as a comprehensive reference for TKGQA and to stimulate further research.

1 Introduction

Knowledge Base Question Answering (KBQA) aims to answer natural language questions based on existing Knowledge Bases (KBs) (Dong et al., 2015). It has garnered significant attention from academia and industry due to its crucial role in various intelligent applications across multiple fields (Zhou et al., 2018). A crucial subtask within KBQA is Temporal Knowledge Graph Question Answering (TKGQA), which specifically addresses temporal questions using Temporal Knowledge Graphs (TKGs) (Leblay and Chekol, 2018a). Temporal questions include temporal constraints or require timestamped answers, reflecting the dynamic and evolving nature of real-world events. The answer can vary significantly with different time constraints. For example, the answer to "Who won the UFC's strawweight championship in 2022?" is "Carla Esparza", while the answer to "Who won the UFC's strawweight championship in 2024?" is "Weili Zhang". Existing KBQA methods, even for complex questions, struggle with temporal questions (Jia et al., 2018b; Sun et al., 2019; Pramanik et al., 2021; Bast and Haussmann, 2015; Abujabal et al., 2017).

Despite growing interest in TKGQA (Chen et al., 2024; Gao et al., 2024; Du et al., 2024; Huang et al., 2024; Xue et al., 2024), the field still grapples with several challenges: (1) Ambiguities in the classification of temporal questions. As illustrated in Table 1, existing methods vary in their understanding of temporal questions, often concentrating on specific types of questions. Currently, there remains an absence of a comprehensive review encompassing all existing temporal questions. (2) Lack of systematic categorization of existing methods. Existing surveys primarily focus on static factual questions and their related KBQA methods (Fu et al., 2020; Lan et al., 2021; Gu et al., 2022; Chakraborty et al., 2021). Considering TKGQA's special handling of timing, it is crucial to conduct an exhaustive review of TKGQA methods.

To address the above challenges, this paper provides a thorough survey from two perspectives: the taxonomy of temporal questions and the methods categorization for TKGQA. Specifically, we first establish a unified taxonomy that encompasses existing temporal question types and definitions, providing a standardized reference that could be widely adopted. Subsequently, we systematically categorize existing methods into semantic parsing-based and TKG embedding-based. Within each category, we highlight how they uniquely address the temporal questions. We identify the temporal question types that each method can solve and summarize them in a table to analyze the focus of existing methods and the question types that lack

Dataset	KG/TKG	Representation Form	Question Types
Temp Questions	Freebase	CVT	Explicit Implicit Ordinal Temp.Answer
Time Questions	Wikidata	triples n-array tuple (n>3)	Explicit Implicit Ordinal Temp.Answer
Crom Questions	Wikidata	quintuple	SimpleTime SimpleEntity Before/After First/Last TimeJoin
MultiTQ	ICEWS05-15	quadruple	Equal Before/After First/Last Equal Multi Before Last After First

Table 1: TKGQA datasets, as well as their background temporal knowledge graphs, the representation form of temporal fact therein, and question types.

attention. Building on this review, we further analyzed the future research directions. To the best of our knowledge, this is the first comprehensive survey on the TKGQA task. This work aims to stimulate further research and foster innovation in the field by serving as a comprehensive reference for TKGQA.

The rest of this paper is organized as follows. In §2, we define in detail the relevant concepts of TKGQA and this task itself. In §3, we classify temporal questions across all datasets based on question content (§3.1), answer type(§3.2), and complexity (§3.3). In §4, we introduce the two categories to TKGQA methods; in §4.1, we detail semantic parsing-based methods, while in §4.2, we elaborate on TKG embedding-based methods; in §4.3, we align each method with the specific types of questions it is designed to solve, providing a detailed table for summary. In §5, we explore new frontiers, summarize their challenges, and highlight opportunities for further research. We conclude this survey in §6. Additionally, in Appendix A, we provide a detailed description of the existing TKGQA datasets (§A.1), including the knowledge graphs behind them; introduce the evaluation metrics (§A.2) for the TKGQA tasks, and provide a leaderboard to illustrate the latest research progress (§A.3).

2 Preliminary

Temporal Knowledge Graph. A TKG usually is denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{F})$, where $\mathcal{E}$, $\mathcal{R}$, $\mathcal{T}$, and $\mathcal{F}$ represent the entities, relations, timestamps, and facts respectively (Cai et al., 2024). A temporal fact $f \in \mathcal{F}$ comprises one or more entities, relations, and associated timestamps. It can be represented in various forms, including Compound Value Types (CVTs), triples, n-array tuples, quintuples, and quadruples.

Temporal Question. A temporal question contains at least one temporal constraint or requires timestamps as its answer (Jia et al., 2018a). A temporal constraint involves a combination of a temporal expression and a temporal word, setting a condition about a specific time point or interval that the answer must meet (e.g., "in 1996"). Temporal expressions refer to time points or intervals with varying levels of granularity in natural language (e.g., "May 11th, 2024") (Pustejovsky et al.; Huang, 2018). Temporal words indicate the temporal relationships between temporal expressions and act as trigger words that impose constraints on the answers (e.g., "in", "after", or "during").

Temporal Knowledge Graph Question Answering. Given the temporal knowledge graph $\mathcal{G}$ and a temporal question q in natural language, the TKGQA task aims to answer the q using either a set of entities $\{e|e \in \mathcal{E}\}$ or timestamps $\{\tau|\tau \in \mathcal{T}\}$ from $\mathcal{G}$.

3 Taxonomy of Temporal Questions

We categorize the questions based on three aspects as illustrated in Figure 1: 1) Question Content: We use several time-related dimensions in question content to categorize the questions, as these dimensions naturally differentiate how questions are answered. 2) Answer Type: We classify the questions based on the answer types; unlike KBQA questions with a single answer type (i.e., entity), temporal questions encompass various types of answers. 3) Complexity: Similar to KBQA, we categorize the questions by their complexity (Hu et al., 2018; Luo et al., 2018).

3.1 Question Content

Temporal Granularity. Questions can be categorized by the temporal granularity of their temporal expressions, with **"year"** being the most common, followed by **"day"** and **"month"**.
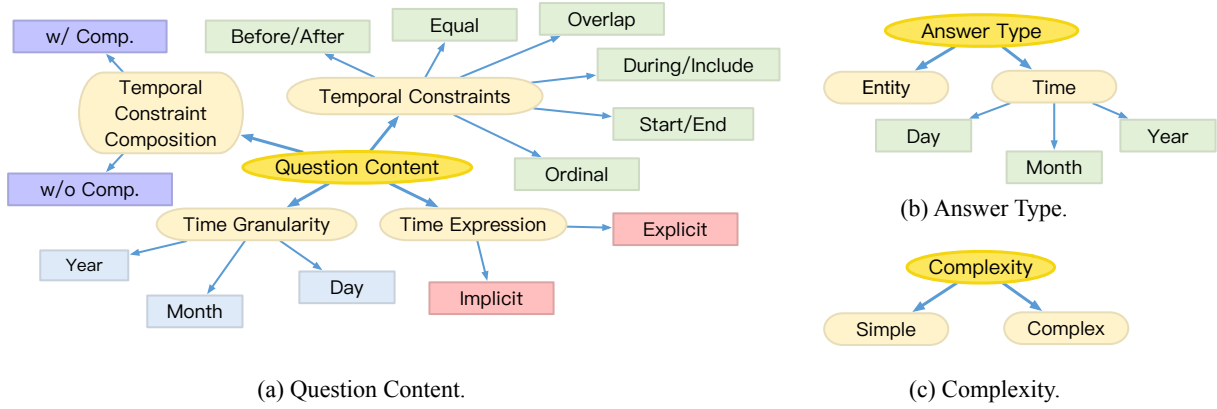
Figure 1: Taxonomy of temporal questions from three aspects, including (a) Question Content; (b) Answer Type and (c) Complexity.

Temporal Expression. Questions can be classified as **explicit** or **implicit** based on the nature of their temporal expressions. All time points can be normalized to a standard format, such as 2024-08-09. Explicit temporal expression can be normalized without additional context (e.g., "September 2023" as 2023-09). Implicit temporal expression, such as an event name or phrase with a temporal scope (e.g., "2024 Paris Olympics"), requires contextual information to be normalized into a specific interval (Jia et al., 2018a).

Temporal Constraints. The types of temporal constraints mirror those of temporal relations between temporal expressions. We simplify Allen's internal algebra for temporal reasoning (Allen, 1983) into six types of relations: *Before/After*, *Equal*, *Overlap*, *During/Include*, *Start/End*, *Ordinal*. Their formalizations are as follows:

- $[begin_{ans}, end_{ans}]$: This represents the time interval or specific time point where the answer is located.

- $[begin_{cons}, end_{cons}]$: This denotes the range of the temporal constraint. When $begin_{cons} = end_{cons}$, it signifies a specific point in time.

A summary of the meanings of these temporal constraint types is provided in Table 2.

The *Ordinal* type requires facts to be arranged in chronological order.

Temporal Constraints Composition. Temporal constraints composition occurs when multiple temporal constraints are in one question. For instance, "Who was the first to request a meeting with Togo in 2005?" combines an *Equal* type constraint "in 2005" with an *Ordinal* type constraint "first". The

Constraint Type	Formalization
Before	$end_{ans} \leq begin_{cons}$
After	$begin_{ans} \geq end_{cons}$
Equal	$begin_{ans} = begin_{cons}, end_{ans} = end_{cons}$
Overlap	$begin_{ans} \leq end_{cons} \leq end_{ans}$ or $begin_{ans} \leq begin_{cons} \leq end_{ans}$
During	$begin_{cons} \leq begin_{ans} \leq end_{ans} \leq end_{cons}$
Include	$begin_{ans} \leq begin_{cons} \leq end_{cons} \leq end_{ans}$
End	$begin_{cons} \leq begin_{ans} \leq end_{cons} = end_{ans}$
Start	$bengin_{ans} = begin_{cons} \leq end_{ans} \leq end_{cons}$

Table 2: Formalization of constraint types.

answer must satisfy both. This combination represents a more complex and challenging type of question.

3.2 Answer type

Temporal questions can require answers that are either collections of **entities** or collections of **timestamps**, with the granularity of the timestamps varying based on the specific question. The type of answer is guided by the question word—such as "who" for entity and "what year" for timestamp.

3.3 Complexity

KBQA works define complex questions as those requiring retrieval of answers from more than one fact (Hu et al., 2018; Dubey et al., 2019). Inspired by these works, we also categorize temporal questions based on complexity. Specifically, we classify temporal questions into simple and complex categories.

Simple questions. Simple questions rely on a single fact for resolution. For instance, "What currency was used in Germany in 2012?" requires retrieving only one fact <Germany, currency, Euro, 2012>.

Complex questions. Complex questions require the integration of multiple facts. For example, the question "Who was the US President before Obama?" first establishes the time constraint "before 2009" based on the fact `<Obama, President of, USA, 2009, 2017>`. The system then identifies the individual who served immediately prior, confirmed by the fact `<George W. Bush, President of, USA, 2001, 2009>`, thus identifying `George W. Bush`. This multi-step reasoning process illustrates the complexity of such questions.

4 Two Categories of TKGQA Methods

Since TKGQA is a crucial subtask within KBQA, many TKGQA methods have been developed to enrich and improve upon KBQA approaches. KBQA methods are categorized into Semantic Parsing-based (SP-based) and Information Retrieval-based (IR-based) methods by existing surveys (Fu et al., 2020; Lan et al., 2021, 2022). Building on this categorization, we classify TKGQA methods into **Semantic Parsing-based (SP-based)** and **TKG Embedding-based (TKGE-based)** methods. Slightly different from IR-based methods in KBQA, TKGE-based methods view TKGQA as a TKG completion task (Cai et al., 2023; Leblay and Chekol, 2018b; Han et al., 2021) and do not always retrieve a question subgraph as in IR-based methods. The following sections delve into the details of these two categorizations of TKGQA methods.

4.1 Semantic Parsing-based Methods

As illustrated in Figure 2, SP-based methods usually have four steps: question understanding, logical parsing, TKG grounding, and query execution. The question understanding module converts unstructured text into encoded questions, facilitating downstream parsing. Next, the logical parsing module transforms the encoded question into uninstantiated logical forms, which are then grounded with the TKG elements through TKG grounding to get executable queries. Finally, the executable queries are processed and executed against the TKG to obtain the final answers during the query execution phase.

4.1.1 Question Understanding

The question understanding module analyzes the input question to generate an encoded representation. This module is sometimes simplified to tag or extract logical candidates like temporal words,
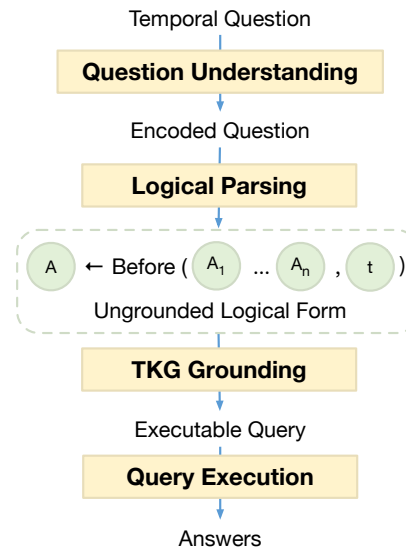

Figure 2: Overall procedure of SP-based methods.

entities, and timestamps. Abstract Meaning Representation (AMR) (Kapanipathi et al., 2020) is one of the most widely used representations for KBQA questions, SYGMA (Neelam et al., 2021) uses AMR to capture temporal words as part of the *:time* relation and handling implicit temporal constraints. Kannen et al. (2023) and Long et al. (2022) also employ AMR to identify question constituents. SF-TQA (Ding et al., 2023) fine-tunes BERT (Devlin et al., 2019) to annotate elements determined by TimeML (Pustejovsky et al.) relations. With its impressive performance on text generation and induction, Large Language Model (LLM) have been applied to generate a simplified version of logical forms directly (Chen et al., 2024) and induce step-wise abstract methodological guidance to the present question (Chen et al., 2023a).

4.1.2 Logical Parsing

Logical parsing transforms the encoded question into an uninstantiated logical form. TEQUILA uses the existing KBQA engines AQQU (Bast and Haussmann, 2015) and QUINT (Abujabal et al., 2017) to answer the sub-questions; these engines primarily rely on predefined rules or templates to parse questions and derive logical forms (Fu et al., 2020). Early TKGQA approaches also employed rule-based translation, further incorporating time-related rules. SYGMA introduces KB-agnostic rules into λ-expressions (Cai and Yates, 2013) to match temporal constraints indicated by the *:time* relation in AMR. Built on SYGMA, Kannen et al. (2023) decompose the λ-expression into main-λ

4

Semantic framework of temporal constraints

Which movie did Alfred Hitchcock [Event1direct] [Signal1in] [Time11960]?

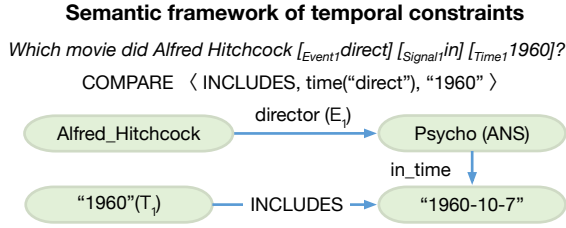COMPARE ⟨ INCLUDES, time("direct"), "1960" ⟩

Figure 3: Semantic framework of temporal constraints.

and aux-λ, with the former containing the primary event questioned and the latter containing the temporal constraint.

Additionally, many methods design specialized logical forms to represent temporal information (Long et al., 2022). Ding et al. (2023) introduce the Semantic Framework of Temporal Constraints (SF-TCons), which captures temporal constraints and their interpretation structures. Six interpretation structures (IS) are summarized based on the intrinsic connection between events and their connectors. For example, the IS-1 Comparison structure 'COMPARE⟨ INCLUDES, time("direct"), "1960" ⟩' in Figure 3 interprets that the "direct" event's time should be "INCLUDES" by "1960". After linking, it can be transformed into the query graph under it. Prog-TQA expands temporal operators based on Knowledge-oriented Programming Language (KoPL) (Cao et al., 2022), which enables a more concise implementation of temporal logical queries compared to KBQA logical forms such as SPARQL (Polleres). ARI defines specialized actions for precise information retrieval, such as "get-Between(entities,Time1,Time2)", which identifies entities/events that occurred between two specific times. An action sequence generated by LLM can be viewed as a logical form here.

4.1.3 TKG Grounding

TKG grounding grounds the elements in the unbound logical form with the entities, relations, and timestamps in the TKG. A series of methods are employed in this module, including rule-based approaches (Neelam et al., 2021), BERT representation similarity (Yih et al., 2015), fuzzy matching algorithms (Chen et al., 2024), and an off-the-shelf Named Entity Linking (NEL) model (Chen et al., 2023a).

4.1.4 Query Execution

The query execution module runs the grounded logical form against the TKG to retrieve the final an-swers. Some methods conduct temporal reasoning during this module. TEQUILA casts sub-questions answers' time range into intervals and conducts reasoning based on rules in Table 2. AE-TQ conducts temporal reasoning using semantic information structures (SISs). One that contains the temporal information computes a temporal constraint, which is then used to filter the candidate answers retrieved by another SIS. ARI performs knowledge-based interaction for multi-step inference (Gu and Su, 2022). The LLM generates and executes actions on the TKG iteratively until the final state provides the answer. Other methods try to enhance model robustness by generating multiple queries: SF-TQA generates multiple candidate queries and scores the pairs of input questions and serialized queries with BERT. Prog-TQA identifies potential errors in KoPL programs and generates corrected versions. Correct programs are collected and used to fine-tune the LLM for self-improvement (Huang et al., 2022) iteratively.

To mitigate the TKG's incompleteness, Kannen et al. (2023) propose a targeted temporal fact extraction technique. Where they use a reading comprehension question answering (RCQA) style model to obtain missing facts and complete the query.

4.2 TKG Embedding-based Methods

As illustrated in Figure 4, TKGE-based methods typically involve three steps: TKG embedding, question embedding, and answer ranking. In these methods, questions and candidate answers (i.e., entities and timestamps) are converted into embeddings through the question embedding and TKG embedding modules, respectively. The question embedding is then projected into Q_{ent} and Q_{time} for ranking entities and timestamps during the answer ranking process.
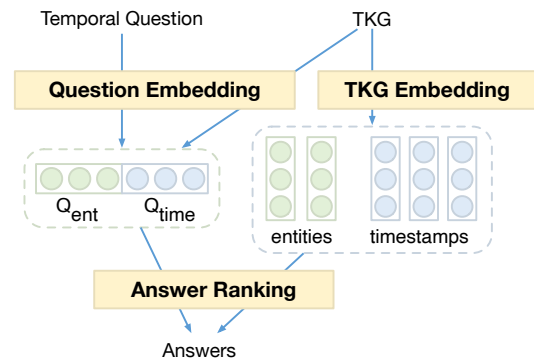

Figure 4: Overall procedure of TKGE-based methods.

4.2.1 TKG Embedding

The TKG Embedding module generates embeddings of TKG elements. The entity and timestamp embeddings are filtered and augmented to create a pool of candidate answers. EXAQT (Jia et al., 2021) follows a line of KBQA research (Sun et al., 2018a; Yasunaga et al., 2022), employing relational graph convolutional networks (R-GCNs) to update and derive the candidates' embeddings. The entity embeddings are initialised with Wikipedia2Vec (Yamada et al., 2020) and argumented with timestamp encodings (Zhang et al., 2020a), time-aware entity embeddings, temporal signals (Setzer, 2001a), temporal question categories (Jia et al., 2018b) and attention over temporal relations.

CRONKGQA (Saxena et al., 2021) initially encodes all elements of the TKG using the TComplEx model (Lacroix et al., 2020), a tensor factorization model designed for temporal knowledge graph completion (Cai et al., 2023), capturing complex patterns and temporal dependencies within multi-relational data. TSQA (Shang et al., 2022a) highlight that TComplEx ignores the temporal order between quadruples; they incorporate temporal order loss during the training of TComplEx, inspired by position embeddings in transformers (Vaswani et al., 2023).

To reduce the search space, EXAQT generates compact question subgraphs using Group Steiner Trees (GSTs) (Li et al., 2016). SubGTR (Chen et al., 2022) crops question subgraphs using temporal constraints.

To address the inconsistency between a question's granularity and the TKG's temporal granularity, MultiQA (Chen et al., 2023b) employs multi-granularity temporal aggregation. It splices days within each month or year interval, adds position vectors, and then fuses the information using the transformer.

4.2.2 Question Embedding

The question embedding module embeds the temporal question, analyzing its semantics and incorporating time-relevant information. EXAQT embeds the question words with Wikipedia2Vec (Yamada et al., 2020) and encodes it with LSTM (Hochreiter and Schmidhuber, 1997). It then concatenates it with temporal category and temporal signal word encodings and updates using R-GCN. Saxena et al. (2021) encodes the question with BERT. TempoQR (Mavromatis et al., 2021) further leverages TKG embeddings to ground questions with their specific entities and respective time scopes. It replaces the BERT token embeddings of entities and timestamps with their pre-trained TKG embeddings and adds time position to the entity tokens. TSIQA (Xiao et al., 2022) derives the time position of entities based on the assumption that entities with co-sharing relations correspond to related timestamps.

Many methods use GNN to further integrate the graphical structure into question embedding; the value of an edge in the graph is the concatenation of relation and timestamp, i.e., $r||t$, which is specific to TKGQA tasks. TwiRGCN (Sharma et al., 2022) computes question-dependent edge weights to modulate R-GCN messages, enhancing messages through relevant edges and diminishing those from irrelevant ones. LGQA (Liu et al., 2023b) fuses global (i.e., sentence-level semantic) and local (i.e., entity-level graphical) information with transformers. GenTKGQA (Gao et al., 2024) retrieves a question-relevant subgraph through LLM's extraction ability (Sun et al., 2023) and uses a pre-trained T-GNN layer (Veličković et al., 2018) to embed elements in the subgraph into "virtual knowledge indicators" to represent question. $M3$TQA (Zha et al., 2024) designs a multi-stage aggregation module, enabling asynchronous alignment and fusion of bidirectional heterogeneous information from the PLMs (Devlin et al., 2019; Liu et al., 2019) and GNNs.

To emphasize the importance of different knowledge for the question, JMFRN (Huang et al., 2024) aggregates entity and timestamp information of retrieved facts using time-aware and entity-aware attention (Vaswani et al., 2023). TMA (Liu et al., 2023a) selects facts with similar semantics for three kinds of token-level attention. A gating mechanism integrates these representations to enhance the question embedding.

To enhance the model's sensitivity to temporal words, TSQA and TSIQA alter temporal words (e.g., replacing "before" with "after") to construct contrastive questions and apply both order loss and answer loss for contrastive learning.

Various approaches extract implicit temporal features from questions: CTRN (Jiao et al., 2023) uses multi-head self-attention, GCN (Sun et al., 2018b), and CNN (Pota et al., 2020) to capture these features and fuse them with augmented BERT representations, while SERQA (Du et al., 2024) integrates temporal constraint features computed from

	Category	Question Content												Answer Type				Complexity	
		Time Granularity			Time Expression		Temporal Constraint						Temporal Constraints Composition		Entity	Time			Simple	Complex
Method		Year	Month	Day	Explicit	Implicit	Overlap	Before/After	Ordinal	Equal	During/Include	Start/End	w/ Comp.	w/o Comp.		Year	Month	Day		
Semantic Parsing-based																				
TEQUILA (Jia et al., 2018b)		○	○	○	○	○	○	○	○	○	○	○		○	○	○	○	○	○	○
SYGMA (Neelam et al., 2021)		○	○	○	○	○	●	○	●	○	●	○		○	○	○	○	○	○	○
AE-TQ (Long et al., 2022)		○			○	○	●	●	●	○	○	○		○	○	○			○	●
SF-TQA (Ding et al., 2023)		○	○	○	●	●	●	●	●	●	○	○	○	○	○	○	○	○	○	●
ARI (Chen et al., 2023a)		○	○	○	○	○	○	○	○	●	○	○		○	○	○	○	○	○	●
Best of Both (Kannen et al., 2023)		○	○	○	○	●	○	○	○	○	○	○		○	○	○	○	○	○	○
Prog-TQA (Chen et al., 2024)		●	●	●	○	○	●	●	●	●	●	●	○	○	○	●	●	●	○	●
TKG Embedding-based																				
CronKGQA (Saxena et al., 2021)		○			○	○	○	○	○	○	○	○		○	○	○			●	○
EXAQT (Jia et al., 2021)		○			●	○	○	○	○	○	○	○		○	○	●			○	○
TempoQR (Mavromatis et al., 2021)		○			○	○	●	●	○	●	○	○		○	○	○			○	●
TSQA (Shang et al., 2022b)		○			○	○	○	●	○	●	○	○		○	○	○			○	●
CTRN (Jiao et al., 2023)		○			○	○	●	●	○	●	○	○		○	○	○			○	●
SubGTR (Chen et al., 2022)		○			○	●	●	●	○	●	○	○		○	○	○			○	●
TwiRGCN (Sharma et al., 2022)		○			○	○	●	●	○	●	○	○		○	○	○			○	○
TSIQA (Xiao et al., 2022)		○			○	○	○	●	○	●	●	○		○	○	○			○	●
TMA (Liu et al., 2023a)		○			○	○	●	●	○	●	○	○		○	○	○			○	○
MultiQA (Chen et al., 2023b)		●	●	●	○	○	○	○	○	○	○	○		○	○	●	●	●	○	○
LGQA (Liu et al., 2023b)		○	○	○	○	○	○	●	○	●	○	○		○	○	○	○	○	○	●
JMFRN (Huang et al., 2024)		○			○	●	○	○	○	●	○	○		○	○	○			○	●
SERQA (Du et al., 2024)		○	○	○	○	○	●	●	○	●	○	○		○	○	○	○	○	○	●
QC-MHM (Xue et al., 2024)		○			●	●	●	●	○	●	○	○		○	○	●			○	●
GenTKGQA (Gao et al., 2024)		○			○	○	●	●	○	●	○	○		○	○	●			●	●
M3TQA (Zha et al., 2024)		○	○	○	●	●	○	○	○	○	○	○	○	○	○	●	●	●	○	●

Table 3: Question category coverage comparison across TKGQA methods. The ○ indicates that this method can solve the corresponding question category. The ● indicates that this method focuses on or specializes in solving this question category.

syntactic information in constituent and dependency trees (Sun et al., 2022; Zhang et al., 2020b; Wang et al., 2023; Liang et al., 2022) combined with Masked Self-Attention (MSA).

To enhance the interpretability of reasoning on implicit temporal questions, SubGTR designs an implicit expression parsing module to rewrite their temporal constraints explicitly.

4.2.3 Answer Ranking

The answer ranking module ranks candidate answers based on the question and candidate answer embeddings. TKG models employ various techniques: leveraging TComplEx scoring functions (Saxena et al., 2021; Mavromatis et al., 2021), applying temporal activation functions to satisfy time constraints (Chen et al., 2022), introducing gating mechanisms (Sharma et al., 2022) or type discrimination losses (Huang et al., 2024) to distinguish among answer types, and fine-tuning a LLM to list the most relevant answers (Ye et al., 2023).

4.3 Question Category Coverage Comparison Across TKGQA Methods

Building on the question taxonomy and methodologies overview, we match each type of temporal question with the appropriate method designed to address it effectively, providing a detailed table

as Table 3. We can see from the table that finer-grained granularities have been brought into focus over time. *Implicit* questions received more attention than *explicit*; *before/after* and *ordinal* questions received the most attention, followed by *during/include* and *overlap*; *start/end* and *equal* questions gain less attention because fewer datasets present them as separate categories. More methods focused on solving more *complex* questions; however, there was a lack of attention to the most complex type of *temporal constraint compositions*.

5 Future Directions

This section will discuss emerging frontiers for TKGQA, aiming to stimulate further research in this field.

5.1 Introduce More Question Types

While existing datasets already cover some of the temporal questions, there are still more questions to be explored in the real world. 1) More combination of existing question types: "Who was the **first** person to win a medal **during** the 2024 Olympic Games?" 2) More time granularity: Some questions demand more fine-grained granularities, such as "When was the Long March 1 launched?" 3) Questions must consider the posed time: "Where

are the seneca indians now?" (Jia et al., 2021; Liška et al., 2022) 3) Predicting the future questions: "Will the Palestinian-Israeli conflict end next year?" (Jin et al., 2021; Ding et al., 2022b,a) 4) Common sense temporal questions: "How often are the Olympics held?"

5.2 Enhance Model Robustness

Most existing TKGQA datasets provide entity and temporal annotations (Saxena et al., 2021; Jia et al., 2021; Neelam et al., 2022), greatly reducing the task's difficulty. Results on unlabeled datasets rely on the effects of NEL or temporal annotators (Chen et al., 2023b), corrupting the model's robustness. Robust models should be able to perform well on datasets with no additional annotations and be able to generalize to unseen entities and relationships (Chen et al., 2022). In addition, most existing datasets rely on template generation and lack diversity; there are very few event types, and they are still single-domain. These can be improved in future work.

5.3 Multi-modal TKGQA

Current TKGQA systems mainly handle plain text input. However, we experience the world with multiple modalities (e.g., language and image). Therefore, building a multi-modal TKGQA system that can handle multiple modalities is an important direction to investigate (Yu et al., 2023). A non-trivial challenge is how to effectively make a multimodal feature alignment and complementary to understand the temporal part better.

5.4 LLM for TKGQA

Recently, Large Language Models (LLMs) have gained significant attention for their remarkable performance across a wide range of Natural Language Processing (NLP) tasks (Touvron et al.; OpenAI, 2024; Team and Googlba, 2024). Existing research has also explored applying LLMs in KBQA scenarios, employing both few-shot and zero-shot learning paradigms (Nie et al., 2024; Sun et al., 2024; Jiang et al., 2023; Baek et al., 2023; Li et al., 2023a,b)

However, several critical challenges remain to be addressed in LLM for TKGQA. We summarize the main challenges as follows: LLMs currently have significant shortcomings in understanding temporal expressions (Chu et al., 2023), crucial for TKGQA. LLMs also perform poorly in symbolic temporal reasoning, especially in multi-step

tasks (Chu et al., 2023; Tan et al., 2023; Qin et al., 2021). Enhancing these capabilities for complex temporal questions is essential; approaches like temporal span extraction pre-training, supervised fine-tuning, and time-sensitive reinforcement learning may help (Tan et al., 2023).

Several emerging opportunities could further enhance the capabilities of LLMs in TKGQA systems:

- **Multi-Agent Collaboration Interactive Reasoning for TKGQA.** Recent LLM works have shifted the focus from traditional NLP tasks to exploring language agents in simulation environments that mimic real-world scenarios (Zhang et al., 2024). Qian et al. (2024) investigates interactive reasoning and collective intelligence in autonomously solving complex problems. This may be further explored for temporal reasoning in temporal questions.

- **Diverse Data Generation.** Numerous studies have demonstrated the effectiveness of large models in data generation (Chung et al., 2023), which can be used to enhance the diversity of the TKGQA dataset.

- **Supplementing Knowledge.** The language model itself can serve as a TKG as demonstrated by Dhingra et al. (2022). Additionally, LLMs possess temporal commonsense (Chu et al., 2023), which is often absent in traditional temporal knowledge graphs. This temporal knowledge can complement existing TKGs for TKGQA.

6 Conclusion

In this paper, we provided an in-depth analysis of the emerging field of TKGQA with a new taxonomy of temporal questions and a systematic categorization of existing methods. We demonstrated the focus and neglect of existing methods for temporal questions, indicating future research directions. We have discussed some new trends in this research field, hoping to attract more breakthroughs in future research.

Limitations

This study offers a comprehensive review of the TKGQA task. However, our primary focus is on temporal question answering specifically based on temporal knowledge graphs, and we do not delve into other temporal question answering tasks based

on texts or heterogeneous sources. Furthermore, the descriptions within this survey are deliberately brief to ensure a broad coverage of the topic while adhering to page constraints. Rather than presenting the works in an unstructured sequence, we organize them into meaningful, structured groups. We aim for this work to serve as an index, guiding readers to more detailed information in the referenced works.

References

Abdalghani Abujabal, Mohamed Yahya, Mirek Riedewald, and Gerhard Weikum. 2017. Automated template generation for question answering over knowledge graphs. In *Proceedings of the 26th international conference on world wide web*, pages 1191–1200.

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. *Preprint*, arxiv:2306.04136.

Hannah Bast and Elmar Haussmann. 2015. More accurate question answering on freebase. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1431–1440.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. Association for Computing Machinery.

Borui Cai, Yong Xiang, Longxiang Gao, He Zhang, Yunfeng Li, and Jianxin Li. 2023. Temporal Knowledge Graph Completion: A Survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6545–6553.

Li Cai, Xin Mao, Yuhao Zhou, Zhaoguang Long, Changxu Wu, and Man Lan. 2024. A Survey on Temporal Knowledge Graph: Representation Learning and Applications. *Preprint*, arxiv:2403.04782.

Qingqing Cai and Alexander Yates. 2013. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 423–433, Sofia, Bulgaria. Association for Computational Linguistics.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base. *Preprint*, arxiv:2007.03875.

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to neural network-based question answering over knowledge graphs. *WIREs Data Mining and Knowledge Discovery*, 11(3):e1389.

Angel X Chang and Christopher D Manning. SUTIME: A Library for Recognizing and Normalizing Time Expressions.

Zhuo Chen, Zhao Zhang, Zixuan Li, Fei Wang, Yutao Zeng, Xiaolong Jin, and Yongjun Xu. 2024. Self-Improvement Programming for Temporal Knowledge Graph Question Answering.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2023a. Temporal Knowledge Question Answering via Abstract Reasoning Induction. https://arxiv.org/abs/2311.09149v2.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023b. Multi-granularity Temporal Question Answering over Knowledge Graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

Ziyang Chen, Xiang Zhao, Jinzhi Liao, Xinyi Li, and Evangelos Kanoulas. 2022. Temporal knowledge graph question answering via subgraph reasoning. *Knowledge-Based Systems*, 251:109134.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models. *Preprint*, arxiv:2311.17667.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Preprint*, arxiv:1810.04805.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Wentao Ding, Hao Chen, Huayu Li, and Yuzhong Qu. 2023. Semantic Framework based Query Generation for Temporal Question Answering over Knowledge Graphs. *Preprint*, arxiv:2210.04490.

Zifeng Ding, Zongyue Li, Ruoxia Qi, Jingpei Wu, Bailan He, Yunpu Ma, Zhao Meng, Shuo Chen, Ruotong Liao, Zhen Han, and Volker Tresp. 2022a. ForecastTKGQuestions: A Benchmark for Temporal

Question Answering and Forecasting over Temporal Knowledge Graphs. *Preprint*, arxiv:2208.06501.

Zifeng Ding, Ruoxia Qi, Zongyue Li, Bailan He, Jingpei Wu, Yunpu Ma, Zhao Meng, Zhen Han, and Volker Tresp. 2022b. *Forecasting Question Answering over Temporal Knowledge Graphs*.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 260–269.

Chenyang Du, Xiaoge Li, and Zhongyang Li. 2024. Semantic-enhanced reasoning question answering over temporal knowledge graphs. *Journal of Intelligent Information Systems*.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *The Semantic Web – ISWC 2019*, pages 69–78, Cham. Springer International Publishing.

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges. *Preprint*, arxiv:2007.13069.

Yifu Gao, Linbo Qiao, Zhigang Kan, Zhihua Wen, Yongquan He, and Dongsheng Li. 2024. Two-stage Generative Question Answering on Temporal Knowledge Graph Using Large Language Models. *Preprint*, arxiv:2402.16568.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. 2018. Learning Sequence Encoders for Temporal Knowledge Graph Completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4816–4821, Brussels, Belgium. Association for Computational Linguistics.

Yu Gu, Vardaan Pahuja, Gong Cheng, and Yu Su. 2022. Knowledge Base Question Answering: A Semantic Parsing Perspective. *Preprint*, arxiv:2209.04994.

Yu Gu and Yu Su. 2022. ArcaneQA: Dynamic Program Induction and Contextualized Encoding for Knowledge Base Question Answering. *Preprint*, arxiv:2204.08109.

Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. 2021. EXPLAINABLE SUBGRAPH REASONING FOR FORE- CASTING ON TEMPORAL KNOWL- EDGE GRAPHS.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Sen Hu, Lei Zou, and Xinbo Zhang. 2018. A State-transition Framework to Answer Complex Questions over Knowledge Base. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2098–2108, Brussels, Belgium. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large Language Models Can Self-Improve. *Preprint*, arxiv:2210.11610.

Rikui Huang, Wei Wei, Xiaoye Qu, Wenfeng Xie, Xianling Mao, and Dangyang Chen. 2024. Joint Multi-Facts Reasoning Network For Complex Temporal Question Answering Over Knowledge Graph. *Preprint*, arxiv:2401.02212.

Ruihong Huang. 2018. Domain-Sensitive Temporal Tagging By Jannik Strötgen, Michael Gertz. *Computational Linguistics*, 44(2):375–377.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1807–1810, Torino Italy. ACM.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex Temporal Question Answering on Knowledge Graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 792–802.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Struct-GPT: A General Framework for Large Language Model to Reason over Structured Data. *Preprint*, arxiv:2305.09645.

Songlin Jiao, Zhenfang Zhu, Wenqing Wu, Zicheng Zuo, Jiangtao Qi, Wenling Wang, Guangyuan Zhang, and Peiyu Liu. 2023. An improving reasoning network for complex question answering over temporal knowledge graphs. *Applied Intelligence*, 53(7):8195–8208.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. *Preprint*, arxiv:2005.00792.

Nithish Kannen, Udit Sharma, Sumit Neelam, Dinesh Khandelwal, Shajith Ikbal, Hima Karanam, and L. Venkata Subramaniam. 2023. Best of Both Worlds: Towards Improving Temporal Knowledge Base Question Answering via Targeted Fact Extraction. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramon Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, and et al.

10

Gliozzo. 2020. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. https://arxiv.org/abs/2012.01707v2.

Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. 2020. Tensor Decompositions for temporal knowledge base completion. *Preprint*, arxiv:2004.926.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. *Preprint*, arxiv:2105.11644.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Complex Knowledge Base Question Answering: A Survey. *Preprint*, arxiv:2108.06688.

Julien Leblay and Melisachew Wudage Chekol. 2018a. Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1771–1776, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Julien Leblay and Melisachew Wudage Chekol. 2018b. Deriving Validity Time in Knowledge Graph. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1771–1776, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Rong-Hua Li, Lu Qin, Jeffrey Xu Yu, and Rui Mao. 2016. Efficient and Progressive Group Steiner Tree Search. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 91–106, New York, NY, USA. Association for Computing Machinery.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023a. Few-shot In-context Learning for Knowledge Base Question Answering. *Preprint*, arxiv:2305.01750.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023b. Unlocking Temporal Question Answering for Large Language Models Using Code Execution. *Preprint*, arxiv:2305.15014.

Shuo Liang, Wei Wei, Xian-Ling Mao, Fei Wang, and Zhiyong He. 2022. BiSyn-GAT+: Bi-Syntax Aware Graph Attention Network for Aspect-based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1835–1848, Dublin, Ireland. Association for Computational Linguistics.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models. *Preprint*, arxiv:2205.11388.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arxiv:1907.11692.

Yonghao Liu, Di Liang, Fang Fang, Sirui Wang, Wei Wu, and Rui Jiang. 2023a. Time-Aware Multiway Adaptive Fusion Network for Temporal Knowledge Graph Question Answering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yonghao Liu, Di Liang, Mengyu Li, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. 2023b. Local and Global: Temporal Question Answering via Information Fusion. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5141–5149, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.

Shaonan Long, Jinzhi Liao, Shiyu Yang, Xiang Zhao, and Xuemin Lin. 2022. Complex Question Answering Over Temporal Knowledge Graphs. In *Web Information Systems Engineering – WISE 2022*, pages 65–80, Cham. Springer International Publishing.

Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge Base Question Answering via Encoding of Complex Query Graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2185–2194, Brussels, Belgium. Association for Computational Linguistics.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Soji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2021. TempoQR: Temporal Question Reasoning over Knowledge Graphs. *Preprint*, arxiv:2112.05785.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G. P. Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, Salim Roukos, Alexander Gray, Guilherme Lima, Ryan Riegel, Francois Luus, and L. Venkata Subramaniam. 2022. A Benchmark for Generalizable and Interpretable Temporal Question Answering over Knowledge Bases. *Preprint*, arxiv:2201.05793.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G. P. Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, Salim Roukos, Alexander Gray, Guilherme LimaRyan Riegel, Francois Luus, and L. Venkata Subramaniam. 2021. SYGMA: System for Generalizable Modular Question Answering OverKnowledge Bases. *Preprint*, arxiv:2109.13430.

11

Zhijie Nie, Richong Zhang, Zhongyuan Wang, and Xudong Liu. 2024. Code-Style In-Context Learning for Knowledge-Based Question Answering. *Preprint*, arxiv:2309.04695.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arxiv:2303.08774.

Axel Polleres. SPARQL. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 1960–1966. Springer.

Marco Pota, Massimo Esposito, Giuseppe De Pietro, and Hamido Fujita. 2020. Best Practices of Convolutional Neural Networks for Question Classification. *Applied Sciences*, 10(14):4710.

Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Uniqorn: unified question answering over rdf knowledge graphs and natural language text. *arXiv preprint arXiv:2108.08614*.

James Pustejovsky, Jose Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling Large-Language-Model-based Multi-Agent Collaboration. *Preprint*, arxiv:2406.07155.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIME-DIAL: Temporal Commonsense Reasoning in Dialog. *Preprint*, arxiv:2106.04571.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question Answering Over Temporal Knowledge Graphs. *Preprint*, arxiv:2106.01515.

A. Setzer. 2001a. Temporal information in newswire articles : An annotation scheme and corpus study.

A. Setzer. 2001b. Temporal information in newswire articles : An annotation scheme and corpus study.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022a. Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs. *Preprint*, arxiv:2203.00255.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022b. Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs. *Preprint*, arxiv:2203.00255.

Aditya Sharma, Apoorv Saxena, Chitrank Gupta, Seyed Mehran Kazemi, Partha Talukdar, and Soumen Chakrabarti. 2022. TwiRGCN: Temporally Weighted Graph Convolution for Question Answering over Temporal Knowledge Graphs. *Preprint*, arxiv:2210.06281.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William W Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. *arXiv preprint arXiv:1904.09537*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018a. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018b. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. *Preprint*, arxiv:2307.07697.

Kailai Sun, Zuchao Li, and Hai Zhao. 2022. Reorder and then Parse, Fast and Accurate Discontinuous Constituency Parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10575–10588, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. *Preprint*, arxiv:2304.09542.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models. *Preprint*, arxiv:2306.08952.

Gemini Team and Googlba. 2024. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arxiv:2312.11805.

Hugo Touvron, Louis Martin, and Kevin Stone. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. *Preprint*, arxiv:1706.03762.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *Preprint*, arxiv:1710.10903.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jiajun Wang, Xiaoge Li, and Xiaochun An. 2023. Modeling multiple latent information graph structures

via graph convolutional network for aspect-based sentiment analysis. *Complex & Intelligent Systems*, 9(4):4003–4014.

Yao Xiao, Guangyou Zhou, and Jin Liu. 2022. Modeling Temporal-Sensitive Information for Complex Question Answering over Knowledge Graphs. In *Natural Language Processing and Chinese Computing*, pages 418–430, Cham. Springer International Publishing.

Chao Xue, Di Liang, Pengfei Wang, and Jing Zhang. 2024. Question Calibration and Multi-Hop Modeling for Temporal Question Answering. *Preprint*, arxiv:2402.13188.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. *Preprint*, arxiv:1812.06280.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2022. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. *Preprint*, arxiv:2104.06378.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2023. Generative Knowledge Graph Construction: A Review. *Preprint*, arxiv:2210.12714.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.

Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. 2023. Prophet: Prompting Large Language Models with Complementary Answer Heuristics for Knowledge-based Visual Question Answering. *Preprint*, arxiv:2303.01903.

Zhiyuan Zha, Pengnian Qi, Xigang Bao, Mengyuan Tian, and Biao Qin. 2024. M3TQA: Multi-View, Multi-Hop and Multi-Stage Reasoning for Temporal Question Answering. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10086–10090.

Xuchao Zhang, Wei Cheng, Bo Zong, Yuncong Chen, Jianwu Xu, Ding Li, and Haifeng Chen. 2020a. Temporal Context-Aware Representation Learning for Question Routing. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 753–761, New York, NY, USA. Association for Computing Machinery.

Yikai Zhang, Siyu Yuan, Caiyu Hu, Kyle Richardson, Yanghua Xiao, and Jiangjie Chen. 2024. TimeArena: Shaping Efficient Multitasking Language Agents in a Time-Aware Simulation. *Preprint*, arxiv:2402.05733.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020b. Fast and Accurate Neural CRF Constituency Parsing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4046–4053.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4623–4629, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

A Appendix

A.1 Datasets

This section introduces the datasets, including their background TKG, size, etc. We provide a question category coverage comparison across TKGQA datasets in Table. 4.

TempQuestions (Jia et al., 2018a) is a benchmark dataset derived from Freebase (Bollacker et al., 2008), where temporal knowledge is stored using compound value types (CVTs). Examples of these CVTs include `footballPlayer.team.joinedOnDate`, `footballPlayer.team.leftOnDate`, `marriage.date`, `amusement_parks.ride.opened`, `amusement_parks.ride.closed`. It includes 1,271 questions with temporal signals, question types, and data sources for testing and evaluation.

TempQA-WD (Neelam et al., 2021) is an adaptation of TempQuestions for Wikidata (Vrandečić and Krötzsch, 2014), fulfilling the need identified by Neelam et al. (2021) for a multi-KG-based dataset to evaluate their model. The dataset comprises:

- 839 questions with corresponding Wikidata SPARQL queries, answers, and categories, along with TempQuestions' information.

- 175 questions with AMR, lambda expressions, entities, relations, and KB-specific lambda expressions, in addition to the above information.

TimeQuestions (Jia et al., 2021) is based on Wikidata, it includes temporal facts of triples, such as `<Malia Obama, date of birth, 04-07-1998>` or maintain more temporal knowledge with qualifiers like `<Barack Obama, position held, President of the US; start date, 20-01-2009; end date, 20-01-2017>`.

Jia et al. (2021) searched through eight KG-QA datasets for time-related questions and mapped them to Wikidata. Questions in each benchmark are tagged for temporal expressions using SU-Time (Chang and Manning) and HeidelTime (Strötgen and Gertz, 2010), and for signal words using a dictionary compiled by Setzer (2001b) and manually tagged with its temporal question category. In total, the TimeQuestions comprises 16,859 questions.

CronQuestions (Chen et al., 2022) utilizes a subset of Wikidata that includes facts annotated with temporal information (Lacroix et al., 2020), such as <Barack Obama, held position, President of USA, 2008, 2016>. Entities extracted from Wikidata with both "start time" and "end time" annotations are transformed into event format (e.g., <WWII, significant event, occurred, 1939, 1945>). The dataset comprises a Temporal KG with 125k entities and 328k facts (including 5k event facts), and 410k natural language questions requiring temporal reasoning.

Complex-CronQuestions Chen et al. (2022) observe that existing benchmarks contain many pseudo-temporal questions. For instance, for the question *"What's the first award Carlo Taverna got?"* there is only one fact related to *Carlo Taverna* in the TKG, which makes the temporal word "first" meaningless as a constraint. They remove all simple and pseudo-temporal questions and filter out questions with less than 5 relevant facts in CronQuestions.

MultiTQ (Chen et al., 2023b) is a dataset derived from ICEWS05-15 (García-Durán et al., 2018), where all facts are standardized as quadruple (s, r, o, t). ICEWS05-15 is notable for its rich semantic information, with a higher average number of relation types per entity than other TKGs. The MultiTQ dataset features several advantages, including its large scale, ample relations, and multiple temporal granularity. ICEWS provides time information at a day granularity, while the authors generate higher granularities, such as year and month, for the questions. MultiTQ contains 500,000 questions, making it a significant resource for temporal question-answering research.

A.2 Evaluation Metrics

Hits@k: This is the most used metric of the TKGQA task. TKGQA method use Hits@1 (accuracy), Hits@3, Hits@5, Hits@10 for evaluation. This metric is set to one if a correct answer appears in the first k positions and zero otherwise.

Precision, Recall and F1 score: This metric is widely used for KBQA task. Precision indicates the ratio of the correct predictions over all the predicted answers. Recall is the ratio of the correct predictions over all the ground truth. F1 score computes the average of precision and recall.

P@1: Precision at the top rank is one if the highest ranked answer is correct and zero otherwise.

MRR: This is the reciprocal of the first rank where we have a correct answer. If the correct answer is not featured in the ranked list, MRR is zero.

Average number of reasoning steps: ARI uses this metric to measure the reasoning steps for each question. The average number of reasoning steps across all tested questions represents this metric.

A.3 Leaderboard

Table 5 presents a leaderboard featuring the top-2 TKGQA models across all mentioned datasets. For semantic parsing-based methods, the widely used metrics are Hits@1 and F1. For TKG embedding-based methods, the commonly used metrics are Hits@1 and Hits@10.

Based on Table 5, we have following observations: (1) Both SP-based and TKGE-based methods are developed to address the TKGQA task, with no clear indication of superiority. According to the Hits@1 result, TKGE-based methods perform better on CronQuestions and TimeQuestions, while SP-based methods excel on MultiTQ. (2) SP-based methods cover more benchmarks than TKGE-based methods. This may be attributed to the flexibility and expressiveness of the logical form, which allows SP-based methods to address a wider range of question types.

Category			Temp Questions	TempQA-WD	Time Questions	Cron Questions	Complex CronQuestions	MultiQA
Question Content	Time Granularity	Year	✓	✓	✓	✓	✓	✓
		Month	✓	✓				✓
		Day	✓	✓	✓			✓
	Time Expression	Explicit	✓	✓	✓	✓	✓	✓
		Implicit	✓	✓	✓	✓	✓	✓
	Temporal Constraint	Overlap	✓	✓	✓	✓	✓	✓
		Equal	✓	✓	✓	✓	✓	✓
		Start/End	✓	✓	✓	✓	✓	
		During/Include	✓	✓	✓	✓	✓	✓
		Before/After	✓	✓	✓	✓	✓	✓
		Ordinal	✓	✓	✓	✓	✓	✓
	Temporal Constraint Composition	w/ Composition						✓
		w/o Composition	✓	✓	✓	✓	✓	✓
Answer Type	Entity			✓	✓	✓	✓	✓
	Time	Year	✓	✓		✓	✓	✓
		Month	✓	✓				✓
		Day	✓	✓	✓			✓
Complexity	Simple			✓	✓	✓		✓
	Complex			✓	✓	✓	✓	✓

Table 4: Question category coverage comparison across TKGQA datasets

Datasets	SP-based			
	Top-1		Top-2	
Metrics	Hits@1	F1	Hits@1	F1
Tempquestions (Jia et al., 2018a)	41.2 (Ding et al., 2023)	41.1 (Ding et al., 2023)	36.2 (Jia et al., 2018b)	37.5 (Jia et al., 2018b)
TempQA-WD (Neelam et al., 2021)	-	41.6 (Kannen et al., 2023)	-	32.0 (Neelam et al., 2021)
TimeQuestions (Jia et al., 2021)	56.5 (Jia et al., 2021)	52.7 (Ding et al., 2023)	53.9 (Ding et al., 2023)	49.9 (Kannen et al., 2023)
CronQuestions (Saxena et al., 2021)	93.7 (Chen et al., 2024)	97.3 (Chen et al., 2024)	70.0 (Chen et al., 2023a)	-
Complex-CronQuestions (Chen et al., 2022)	-	-	-	-
MultiTQ (Chen et al., 2023b)	79.7 (Chen et al., 2024)	91.0 (Chen et al., 2024)	38.0 (Chen et al., 2023a)	-

(a) Leaderboard for TKGQA datasets (SP-based).

Datasets	TKGE-based			
	Top-1		Top-2	
Metrics	Hits@1	Hits@10	Hits@1	Hits@10
Tempquestions (Jia et al., 2018a)	-	-	-	-
TempQA-WD (Neelam et al., 2021)	-	-	-	-
TimeQuestions (Jia et al., 2021)	62.8 (Huang et al., 2024)	-	60.5 (Sharma et al., 2022)	-
CronQuestions (Saxena et al., 2021)	97.8 (Gao et al., 2024)	99.3 (Chen et al., 2022)	97.1 (Xue et al., 2024)	99.2 (Xue et al., 2024; ?)
Complex-CronQuestions (Chen et al., 2022)	92.0 (Chen et al., 2022)	98.6 (Chen et al., 2022)	79.2 (Mavromatis et al., 2021)	95.9 (Mavromatis et al., 2021)
MultiTQ (Chen et al., 2023b)	29.3 (Chen et al., 2023b)	63.5 (Chen et al., 2023b)	-	-

(b) Leaderboard for TKGQA datasets (TKGE-based).

Table 5: Leaderboard for TKGQA datasets.